DOCUMENT RESUME

ED 285 905 TM 870 497

AUTHOR Mandeville, Garrett K.
TITLE On the Identification of Effective and Ineffective Schools.
PUB DATE Apr 87
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Educational Assessment; Educational Research; Elementary Education; *Identification; *Research Design; Research Problems; *School Effectiveness; Scoring; State Surveys
IDENTIFIERS Aggregation (Data); Cognitive Skills Assessment Battery; Comprehensive Tests of Basic Skills; *Effective Schools Research; Residuals (Statistics); South Carolina; South Carolina Basic Skills Assessment Program

ABSTRACT

An alternative method of identifying effective schools, based on the concept of effective schools as statistically atypical, was tested. This paper investigates the issue of consistency of the identification of what is termed "exceptional" schools. The following tests were administered to all first through fourth graders in 431 schools in South Carolina: (1) South Carolina Basic Skills Assessment Program, given to all first, second and third graders; (2) Comprehensive Tests of Basic Skills, given to all fourth graders; and (3) Cognitive Skills Assessment Battery, given to all entering first graders. Data were collected during 1984-85 and 1985-86. Using studentized residuals for reading, mathematics, and reading and mathematics combined, this study tested for both cross-grade and cross-year consistency of scores. The strongest consistencies were found with combined reading-mathematics scores, at each grade level within schools. Because this combined score appeared to be consistent within grades and across years, it was suggested that it might be a more accurate statistic for predicting which schools are effective (positively exceptional) and which are ineffective (negatively exceptional). (JGL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED285905

On the Identification of Effective and Ineffective Schools

Garrett K. Mandeville

University of South Carolina

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. K. Mandeville

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual convention of the American Educational Research Association, Washington D. C., April 1987.

TM 870 497

BEST COPY AVAILABLE

For a number of years, researchers have examined the success of schools in their efforts to foster important educational outcomes in the children who attend them. The recent impetus in this area is related to state- and district-level programs to monitor school performance on the basis of student achievement test data. In some cases, these programs lead to recognition of high performing schools and in a few cases, monetary awards to the schools and/or their personnel (see, e. g., Wynne, 1984). At the state level California, Florida, and South Carolina have school award programs in which test scores are a major factor in the determination of awardees and the same could be said for a number of district-level programs.

This interest has led to a number of research papers dealing with methodological problems in the identification of exceptional schools. Although these papers have rather diverse objectives, they tend to concentrate on comparing the results obtained from the various methodologies which have been proposed. When similar methodologies have been compared (e. g., assorted regression approaches) the results tend to be quite consistent (e. g., Webster and Olson, 1984; Abalos, Jolly, and Johnson, 1985), but when the procedures vary on major dimensions, the reverse is usually true (e.g., Frechtling, 1982; Frederick and Clauset, 1985).

Much of the earlier "effective schools" research was criticized (see, e.g., Rowan, Bossart & Dwyer, 1983; Ralph & Fennessey, 1983) for being based on somewhat limited data (e.g., one grade, one subject, one year). With these criticisms in mind, those charged with the identification of exceptional schools often use methods which involve the computation of indices for each of a number of subject areas at each grade level followed by their aggregation to the school level.

While this is a reasonable approach to solving the problem of "comprehensivity", it begs the issue of the consistency of that performance over subject areas and grade levels to say nothing about the consistency across years.

This paper will investigate the issue of consistency of the identification of what will be termed "exceptional" schools. The nondescript term--exceptional schools--is used for two reasons: 1) the results may have applicability in a variety of settings (e.g., the identification of effective schools or the identification of schools to receive an award), and 2) schools performing unusually poorly will also be included. It is clear that few educators would agree completely on how to operationalize effectiveness. The arguments against using student achievement as the sole criteria upon which to judge the merits of a schools are clearly valid. However, in this paper as in so many others, student achievement will be the basis of exceptionality.

Literature Review

Of the four general approaches to the identification of effective schools outlined by Rowan et al. (1983), those which are most popular and which appear to have the most empirical support are based on regressing achievement onto prior achievement and some measure of school social standing (SES). An early regression-based methodology was developed by Dyer and his colleagues (Dyer, Linn & Patton, 1969; Dyer, 1970). In the "student change model of an educational system" student outcomes are regressed onto student inputs and "hard-to-change" variables and the residuals are used to describe school performance. Dyer's system uses longitudinally matched student data, the justification being that "the only fair index of school effectiveness is one that rests on input-output data obtained only on those pupils with whom the school staff has been in continuous contact over a

specified period of months or years." (Dyer, 1970, p. 208).

Consistency Across Subject Areas

Rowan and Denk (1983) conducted a study using 405 randomly selected California elementary schools. Sixth grade California Assessment Program (CAP) test scores in reading, math and writing were regressed onto 3rd grade test data and two measures reflecting socio-economic-status (SES). Because matrix sampling is used in the CAP testing, school-wide estimates rather than student level test data were used in the analysis. Also, as the authors note, each of the predictors suffers from some form of aggregation bias and/or measurement error. In spite of these technical difficulties, it is likely that the Rowan and Denk results are essentially correct. The correlations of cross-subject residuals for the three subjects were all in the .60s.

Mandeville and Anderson (1986) analyzed over 400 elementary schools using Dyer's methodology as applied to test scores in reading and math for grades 1-4. The reading and math residuals correlated .70 in grade 1 and were in the low .60s for the other three grades. When a residual was defined as "exceptional" if its (standardized) value was 1.0 or more, the Kappa coefficients were .52, .53, .46, and .33 for grades 1-4 respectively.

Matthews, Soder, Ramey and Sanders (1981) studied the stability of residuals from regression analyses using the public schools in Seattle as their sample. They reported that the SEIs were quite unstable across grades (2-8), subject areas (reading and math), and years (1978-79 and 1979-80) but gave no specific results dealing with inconsistencies as a function of subject or grade.

Consistency Across Grades

Mandeville and Anderson (1986) found cross-grade residuals to be

very small but statistically significant in math (median r = .13) but essentially independent in reading (median r = .06). As noted above, Matthews et al. (1981) also indicated that the residuals were quite unstable across grades but gave none of the details.

Consistency Across Years

The data used in the Rowan and Denk (1983) study covered the three-year period from 1975 to 1977. The correlations of the residuals for adjacent years (i.e., 75-76 and 76-77) were very similar and slightly larger than those for the two year (1975-77) span. The (median) adjacent-year results were .25, .30, and .45 in reading, writing, and math, respectively.

The Matthews et al. (1981) study produced year-to-year correlations which ranged from -.24 to .44, none of which were statistically significant because of the small number of schools involved. The authors also noted that in some cases a school was identified as a positive outlier one year and a negative outlier the next.

Forsyth (1973) applied Dyer's methodology to two successive 12th grade classes in a sample of 50 schools and found that the correlations among the residuals for the two years ranged from .11 to .50 (depending on subject area). When he categorized the schools into five groups based on the sizes of the residuals perfect agreement was rare (16 percent to 36 percent).

On Metrics and the General Approach

The studies included in the above discussion have been similar in a number of respects. All have employed a regression approach with the residuals serving as the basis for assessing school "performance". In most cases, only the correlations of the residuals were considered as evidence concerning consistency. Furthermore, when interest centered

on classification decisions, one of two rather arbitrary approaches were used to classify the schools. That is, either some predetermined number or percentage (e.g., the top 5% and bottom 5% of the schools based on the distribution of residuals) or some set criterion (e.g., those schools with standardized residuals larger than 1.0 in absolute value) were considered exceptional.

In an early study on the identification of effective schools, Klitgaard and Hall (1973) presented arguments for a somewhat different approach to this problem. Essentially, these authors conceptualized exceptional schools as atypical and suggested that they should be identifiable as such in some probabilistic or statistical sense. Thus, they employed procedures such as examining the tails of a distribution of residuals for "lumpiness" and comparing the distribution of cumulative measures of effectiveness (e.g., the number of times a school was designated as atypical over a four year period) to a null distribution.

Methods and Data Sources

South Carolina has tested all students in the majority of the grades in the K-12 grade span for a number of years. Relevant to this research is the fact that criterion referenced tests (CRT) administered under the Basic Skills Assessment Program (BSAP) are given each spring to all students in grades 1, 2, and 3; that 4th graders are spring tested with the Comprehensive Tests of Basic Skills (CTBS); and that the Cognitive Skills Assessment Battery (CSAB) is administered each fall to all 1st graders as a readiness test. The BSAP reading and mathematics tests are relatively short containing 36 and 30 multiple choice items respectively.

This study was limited to the elementary schools in South Carolina which contained grades one through four (and possibly

additional grade levels) for both the 1984-85 and 1985-86 school years. Test records of students in these four grades during the Spring 1985 and Spring 1986 testings were matched with their records for the most recent previous testing. Schools with fewer than 20 matched, complete student records for both years at each of the four grade levels were eliminated producing a final sample of 431 schools.

The Creation of Studentized Residuals

Dyer's approach was used to create residuals for each of two subject areas (reading and math), each of four grades (1-4) and each of the two school years 1984-85 and 1985-86. In these analyses BSAP scale scores in reading and mathematics (grades 1-3), and expanded scale scores for the Total Reading and Total Mathematics subtests of the CTBS (4th grade) based on the Spring 1985 and 1986 testings were used as the output variables; the most recent BSAP or CSAB data were considered to be student input variables; and the percentage of children eligible for free or reduced price lunches was used as a hard-to-change variable. Mean test scores at the (grade within) school level and the lunch percentage were obtained by aggregating the student level data files.

Each of the two output variables (for a given grade and year) were regressed onto the same set of input and hard-to-change variables. [In total, 16 regression analyses (2 subjects x 4 grades x 2 years) were conducted.] The residuals were divided by their estimated standard errors to produce what have been called studentized residuals (SRs) which were to serve as the basis for identifying exceptional schools. For the details of the regression analyses, see Mandeville (1987).

"Standard" Analyses

The initial analyses addressed the question of year-to-year decision consistency following methods which are currently commonly employed. One aspect of that "common practice" which makes sense from

a logical and political standpoint is to accumulate the evidence concerning the effectiveness of a school across the various segments of its population. To facilitate this approach, weighted averages of the residuals were obtained for each year across the two subject areas, across the four grades within each subject, and across both subjects and grades.

An arbitrary decision was made to identify approximately the top 10% of the schools in a distribution as exceptionally effective and the bottom 10% as exceptionally ineffective. For the grade and subject area specific SRs, the distributions of which were approximately unit normal, cut-offs of +/- 1.2 were used. The distributions of SRs based on aggregation, however, were somewhat less variable and the following cut-offs were selected: for the SRs based on reading and math averages (for a given grade) 1.1 was used; for the SRs based on either reading or math averaged across the four grades .7 was used; and .6 was used for the overall aggregate SRs (across subject areas and grades). Year-to-year decision consistency was indexed for each of the 15 SRs using coefficient Kappa.

A Statistical Search

At this stage, it seemed appropriate to take an alternative view of the problem of identifying exemplary schools which was more in line with the early work of Klitgaard and Hall (K-H, 1973). Specifically, two aspects of the K-H work will be applied here. First, the distributions of the SRs will be investigated for outliers (K-H used the term lumpiness in the tails). The concept of an effective school suggests that such schools should be easily discernible in the upper tail.

A second approach will consider the issue of decision consistency. This part of the study was conducted to see if the extremely

discouraging results of Mandeville and Anderson (1986) concerning consistency across grades could be refuted using a more refined approach. Consistency across years was also researched. The basic approach followed that of K-H (pp. 28-30).

For these analyses a criterion for identification as exceptional had to be selected. For most of the situations studied, the criterion (CRIT) of 1.0 was selected; in a few cases, smaller values such as .50 and .75 were also used. Schools with SRs greater than CRIT in absolute value were considered as exhibiting positive or negative exemplary performance.

The cross-grade analyses were conducted for reading and math separately and jointly. For the separate analyses, each school was "scored" (at each grade level) against CRIT so as to obtain a +1, 0, or -1 score. The total score S was obtained by summing the grade specific scores; it ranged from +4 to -4. The distribution of S was then compared to a null distribution (based on the assumption of independence across grades) with a chi square test.

Scoring for the joint analyses at each grade involved comparing both SRs to CRIT and awarding scores ranging from +2 (both SRs larger than CRIT) to -2 (both less than -CRIT). Upon aggregation, the values of S ranged from +8 to -8. Again a chi square test of independence across grades was used.

Cross-year analyses were conducted at each of the four grade levels. The +2/-2 scoring scheme was applied to each of the two years yielding S distributions ranging from +4 to -4. Here the chi square test addressed the null hypothesis of independence across years (within each grade level).

Results

Results of "Standard" Analyses

The statistics on decision consistency are presented in Table 1.

Table 1 about here

The Kappa coefficients based on grade-specific SRs in reading ranged from .41 (1st grade) to .17 (3rd grade). Even at the 1st grade, however, the combined percentage of schools which were inconsistently classified as positive exemplary (% +0/0+ = 8.6) was larger than the percentage which were consistently classified thusly (% ++ = 6.0). Examining the negative exemplary classification for 1st grade reading we see that almost 2.6 times as many schools were inconsistently classified as were consistently classified (% -0/0- = 13.0 vs. % -- = 5.1). We might hope that this inconsistency at the lower end of the distribution indicates that schools which performed poorly during the 1984-85 school year took some positive steps to improve. In fact, at each of the four grades, the data suggest slightly more consistency at the positive than negative end of the scale for reading. The cross-grade aggregate in reading (R1-4) produced the least year-to-year consistency with a Kappa of .08. Since the null standard error of Kappa for this situation is slightly less than .04, even this value is statistically significant at traditional Type I error rates. A significant Kappa is little solace, however, when the percentages on decision consistency are inspected. On the other hand, there were very few cases where a school switched from positive to negative exemplary status (or visa-versa) during the two years.

In math Kappa decreased in an approximate linear fashion from .31 (1st grade) to .21 (4th grade) but generally the values were larger than those for reading. Kappa for the cross-grade math aggregate (M1-4) was .32, larger (in the descriptive sense) than any of the grade-

specific values. With the exception of grade 3, where 7 schools (1.6%) changed from positive to negative exemplary status, very few schools changed so radically over the two years when gauged by the math performance of their students.

Kappas for the cross-subject aggregates (R&Ms) were inconsistently related to the subject-specific reading and math values, being larger than both (grade 2), smaller than both (grade 4) or between them. The aggregate across both subjects and grades (R&M1-4) produced a modest Kappa of .22.

The percentages of schools consistently classified for the two years ranged from 68 to 78 (not shown in the table). However, well over 60% of these were nonexceptional middle category schools. With the exception of 1st grade reading, many more schools (from 2 to 9 times as many) were classified into one of the two exceptional categories only one year than were consistently classified both years.

Results of Statistical Search

Data on the shape of the SR distributions is presenteed in Table 2. The reading and math residuals for grades 1 and 2 are modeled

---

Table 2 about here

---

rather well by a unit normal distribution and there is certainly no evidence of "lumpiness" in either tail. Grade 3 and grade 4 reading do not fit the normal model because the distributions are more heavy-tailed than the normal. The grade 3 distributions are reasonably symmetric so that about the same number of positive and negative outliers are identified; the 4th grade distributions are negatively skewed and for the 1985 data it is clear that more "lumpiness" occurs in the lower than the upper tail. The outlier figures for 1986 (13/5)

are a bit of an anomaly, but the distribution is negatively skewed.

The results for the 3rd and 4th grade math SRs are inconsistent year-to-year. The 1986 data are adequately fit by a normal model. For 1985, however, the 3rd grade data are positively skewed and a large number of positive outliers are identified. The situation is reversed for the 4th grade 1986 math data--which exhibits the same characteristics as the 1986 reading data--namely that negative outliers are much more common.

Results of the chi square analyses for the separate R and M analyses (and CRIT=1) are presented in Table 3. Values of S larger

---

Table 3 about here

---

than 3 or less than -3 were combined because of small expected frequencies. Of the four analyses, the only one which is statistically significant at common Type I error rates is the 1985 math distribution (chi square = 37.8 on 6 d.f., p (.001). The main contributor to this statistic is the S (= -3 cell where O=17 vs. E=5.5 leads to a cell chi square of 23.7. The two top categories of S (S)=3 and S=2) also had rather large contributions to the total chi square, both being larger than 5. The 1986 math data fit the null model rather well, however, as do the reading data for both years.

The joint analyses were conducted in the hope that the known association between reading and math performance might produce a somewhat more powerful procedure for demonstrating the existence of outliers. Values of CRIT ranging from .5 to 1.0 were used here to investigate whether evidence of consistently exceptional performance might be stronger if the operational definition of exceptionality were lowered from the commonly employed value of 1.0. For the smaller

values of CRIT, the probabilities of +2 and -2 scores increase and extreme values of S are more likely. Thus, different rules for combining the tail values of S were used for the different values of CRIT.

Partial results are presented in Table 4. Since our interest is mainly in the tails of the distributions, the observed and expected frequencies are provided only for absolute values of S of four or

---

Table 4 about here

---

larger. Although some of the expected frequencies are less than 5, they roughly satisfy various relaxed criteria and the chi squares may be considered valid. A number of interesting inferences may be made from Table 4.

First, the quality of the fit to the null distribution appears to be poorer as the value of CRIT decreases. This suggests that there exist a number of schools which have a rather consistent positive or negative effect on student learning but that this effect is not as strong as we might expect. It is also possible that differential criteria may need to be used for different subject areas. This would be in line with findings which suggest that schools may have a larger impact on math than reading performance (see, e.g., Mandeville & Anderson, 1986).

Second, the observed and expected frequencies in the tails of the S distributions concur with the alternative hypothesis that schools are more consistent than chance (across grades) in either promoting or detracting from student achievement. That is, for the tail categories of S which are contributing to the total chi square, the observed frequencies exceed the expected frequencies. To some extent, this

result conflicts with the correlational evidence cited earlier that indicated very modest associations in math and essentially no relationship in reading.

Third, there is a hint that exceptionality may be more common on the negative than the positive side. In each of the six chi square analyses, the total frequency of the two most extreme negative categories was larger than the sum of the two most extreme positive ones. This may have important implications for identifying schools where improvement programs are most needed.

The final analyses involved cross-year consistency; the within grade results are presented in Table 5. The data are presented for S ranging from +4 to -4, even though the validity of the analysis is borderline considering the sizes of the expected frequencies in the tail categories. However, when a more conservative approach was taken by combining the two extreme values in each tail, there were no substantive changes in the values of chi square (the conservative chi square values are presented at the bottom of the table).

Table 5 about here

The chi squares are all significant beyond the .001 level for each grade. Once again, the contributions from the tails competely dominate the total chi squares. The data provide strong evidence that there are many more schools than would be predicted by chance which have a consistently positive or negative effect on the children at a particular grade level.

A final cross-year analysis was conducted to assess the consistency of the schools in the tails of the S distributions represented in Table 4. Thus, this is an alternative approach to

assessing the year-to-year consistency using data for all subject areas and grade levels. Whether counting or averaging, however, a criterion--or in this case multiple criteria--must be established. In this case the CRIT=.5 criterion for a given SR was used and values of S of five or larger in absolute value were used to signify exceptionality. According to these standards, there were 26 and 25 positive exceptional schools in 1984-85 and 1985-86 respectively. Of these six were exceptional both years. The corresponding frequencies for negative exceptional schools were 41 and 38 for the two school years and 11 were consistently categorized. Although not impressive, the Kappa for these data was .34 which is somewhat larger than the Kappa for the R&M1-4 aggregation approach (.22) presented in Table 1. (The Kappa for CRIT=.75 was only .18.)

Summary

An assumption implicit in attempts to categorize schools on the academic achievement of their pupils is that all segments of that pupil population are being treated equitably, that progress (or lack of it) is being made in a number of the important learning areas and that schools will more-or-less maintain the same relative status from year to year. One way to address the validity of these assumptions is to calculate SRs and correlate them. Based on previous research, SRs correlate in the .50s to .60s across subject areas, in the 0 to .15 range across grades, and between about .10 and .50 across two successive years. Thus, the assumption of cross-grade consistency is of questionable validity and consistency across years is somewhat suspect.

The "standard analyses" conducted in this study were designed to provide more information on the cross-year question but used a decision making rather than a correlational approach. Furthermore, schools were

classified as exemplary in either the positive or negative sense, or nonexceptional (i.e., a three category system was used). The results were: 1) Kappa coefficients were significant but not large; 2) in most cases, many more schools (2 to 9 times as many) were identified as exceptional only one of the two years than were consistently identified as exceptional; 3) there was some evidence that positive exceptionality was more stable in reading than negative exceptionality; and 4) aggregating across grades was disastrous in reading, produced a slightly larger Kappa in math, and an intermediate value (.22) for reading and math combined.

The second phase of the analyses addressed a more basic issue, namely, can schools be identified as exceptional in a statistical sense. Both cross-grade and cross-year consistency were considered.

Initially, the shape of the SR distributions was considered. Of special interest was whether outliers would be prevalent. There was some evidence that the SR distributions are heavy-tailed in 3rd and 4th grade reading (both years) and math (1984-85 only); in some cases there were a number of outliers in both tails but this was not always the case as some of the distributions were skewed (usually negatively). For grades 1 and 2 the SRs fit a normal model quite well. Overall, there was little evidence of lumpiness in the tails of the SR distributions.

The chi square analyses dealt with the efficacy of an alternative way to combine the SRs available for a school. Cross-grade (within year) consistency was considered first and CRIT was set to 1 for the analyses in which reading and math SRs were treated separately. Reading SRs were consistent with a model which posited independence across grades, not surprising based on the earlier results from the standard analysis. For math the results were mixed with the

independence hypothesis rejected for the 1984-85 data but accepted for the 1985-86 data.

Next a model which took advantage of the relationship between reading and math SRs was considered. These joint cross-grade analyses were conducted with varying values of CRIT. For CRIT=.75 and .50, the independence model was soundly rejected for both years, providing rather strong evidence that cross-grade consistency of some type exists.

Cross-year analyses were conducted using the chi square approach separately for each grade. At all grade levels, the results were highly significant, indicating that within grade, the SRs are not independent year-to-year. One of the most interesting results was found when the two groups of extreme schools for the two years (CRIT=.5 in Table 4) were compared resulting in a Kappa of .34. This represents the strongest evidence for year-to-year consistency based on an approach which utilizes the data for both subjects and all grade levels.

Conclusions

It is rather clear that a major component in the consistency of adjusted student achievement is localized in <u>specific grades within schools</u> rather than more globally manifest across all grades. The implications for the identification of schools for study either 1) to determine the positive characteristics with the objective of exporting them to other schools (effective schools) or 2) to assist in school improvement programs (negative exceptional) are obvious.

However, it is likely that the current emphasis on singling out exceptional schools--based on the democratic ideal of using a broad array of test data--will not soon abate. The results presented here indicate that combining these data using a simple aggregation

strategy has little merit and may actually produce completely unreliable school measurements. On the other hand, a disjunctive model based on "scoring" the reading and math residuals at each grade (possibly using different scoring criteria) and accumulating these scores, was found to provide improved year-to-year consistency. Although it is too early to tell, such an approach may 1) satisfy political desires to make decisions regarding exceptionality based on data from all sectors of a school population, and 2) produce measurements which have the requisite psychometric properties.

References

Abalos, J., Jolly, S. J. & Johnson, R. (1985). Statistical methods for selecting merit schools. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Dyer, H. S. (1970). Toward objective criteria of professional accountability in the schools of New York City. Phi Delta Kappan, 52, 206-211.

Dyer, H. S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy scores based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6, 591-605.

Forsyth, R. A. (1973). Some empirical results related to the stability of performance indicators in Dyer's student change model of an educational system. Journal of Educational Measurement, 10, 7-12.

Frechtling, J. A. (1982). Alternative methods for determining effectiveness: Convergence and divergence. Paper presented at the annual meeting of the American Educational Research Association, New York.

Frederick, J. M. & Clauset, K. H. (1985). A comparison of the major algorithms for measuring school effectiveness. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Klitgaard, R. E. & Hall, G. R. (1973). A statistical search for unusually effective schools. Santa Monica, CA: Rand Corporation.

Mandeville, G. K. (1987). The stability of school effectiveness indices across years. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Mandeville, G. K. & Anderson, L. W. (1986). A study of the stability of school effectiveness measures across grades and subject areas. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 27: 497)

Matthews, T. A., Soder, J. B., Ramey, M. C. & Sanders, G. H. (1981). Use of district test scores to compare the academic effectiveness of schools. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Ralph, J. H. & Fennessey, J. (1983). Science or reform: Some questions about the effective schools model. Phi Delta Kappan, 64, 689-702.

Rowan, B., Bossart, S. T., & Dwyer, D. C. (1983). Research on effective schools: A cautionary note. Educational Researcher, 12, 24-31.

Rowan, B. & Denk, C. E. (1983). Modelling the academic performance of schools using longitudinal data: An analysis of school effectiveness measures and school and principal effects on school-level achievement. San Francisco, CA: Far West Laboratory for Educational Research and Development.

Webster, W. J. & Olson, G. H. (1984). An empirical approach to identifying effective schools. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Wynne, E. A. (1984). School award programs: Evaluation as a component of incentive systems. Educational Evaluation and Policy Analysis, 6, 85-93.

Table 1
Statistics on Decision Consistency
For the 1984-85 and 1985-86 School Years
(N = 431 Schools)

	Percentage				N	
SR	++	+0/0+	-0/0-	--	+-/-+	Kappa
Reading						
R1	6.0	8.6	13.0	5.1	0	.41
R2	4.2	13.7	15.1	2.6	1	.19
R3	3.0	13.2	14.6	2.8	2	.17
R4	3.0	11.6	13.0	2.3	4	.20
R1-4	1.6	14.6	15.8	2.3	2	.08
Math						
M1	3.9	9.5	14.2	4.6	1	.31
M2	3.7	14.2	13.2	5.1	1	.27
M3	4.6	10.4	15.1	2.1	7	.24
M4	2.6	11.6	13.7	3.5	0	.21
M1-4	4.2	11.8	13.5	5.6	1	.32
Combined						
R&M1	5.1	10.9	13.2	5.1	1	.35
R&M2	3.7	12.8	12.3	4.6	0	.29
R&M3	3.3	12.8	14.6	2.6	2	.18
R&M4	2.3	12.8	13.0	3.0	0	.18
R&M1-4	2.8	15.1	15.8	5.8	4	.22

Table 2
Statistics Related to the Shape of SR Distributions
For the 1984-85 and 1985-86 School Years

	Reading				Math			
Grade	1	2	3	4	1	2	3	4
Skewness								
1984-85	-.04	.12	.03	-.89	.01	.00	.62	-1.01
1985-86	.05	.13	.43	-.28	.04	.05	.48	-.20
Kurtosis								
1984-85	.29	.47	1.00	3.41	.54	-.06	1.30	2.48
1985-86	.15	.08	1.01	1.22	.01	.57	.69	.52
Normal-pa								
1984-85	NS	NS	.03	<.01	NS	NS	<.01	<.01
1985-86	NS	NS	<.01	.02	NS	NS	NS	NS
Outliers +/-b								
1984-85	1/1	5/4	9/9	7/18	3/2	4/0	17/2	1/13
1985-86	2/1	5/3	5/2	13/5	4/1	3/2	3/0	2/5

Note: Although not shown, the means and st. devs. never differed from 0 and 1 by more than a few hundredths.
a Based on the Kolmogorov D statistic.
b Frequencies beyond 1.5 x inter-quartile range from the 25th or 75th percentiles.

Table 3
Chi Square Results for Cross-grade Analyses
Reading and Math Treated Separately, CRIT=1

	Reading				Math			
	1985		1986		1985		1986	
S	O	E(CHI2)	O	E	O	E(CHI2)	O	E
)=3	4	3.6	6	3.8	7	3.0(5.2)	6	3.3
2	31	27.1	28	27.1	35	23.6(5.5)	33	25.1
1	101	100.0	95	96.6	78	91.0	82	94.4
0	164	168.0	162	161.1	156	162.2	160	164.0
-1	96	101.3	98	104.9	110	110.6	104	107.3
-2	25	27.7	35	31.8	28	35.0	41	32.4
(=-3	10	3.7(10.5)	7	4.8	17	5.5(23.7)	5	4.9
Chi Square	11.8			3.1		37.8***		8.8

Note: CHI2 is the contribution to chi square and is included only if it exceeds 3.0; the d.f. for chi square are 6 and *** indicates p (.001; all other chi squares are not significant at the .05 level.

Table 4
Partial Chi Square Results for Cross-grade Analyses
Reading and Math Treated Jointly, CRIT = .5, .75, and 1.0

	CRIT=.5				CRIT=.75				CRIT=1.0			
	1985		1986		1985		1986		1985		1986	
S	O	E+	O	E+	O	E+	O	E+	O	E+	O	E+
)=7	1	2.1	6	2.0+	-	-	-	-	-	-	-	-
6/)=6	11	5.0+	7	4.8	4	2.1	6	2.3+	-	-	-	-
5/)=5	14	11.2	12	10.8	6	5.5	8	5.7	6	2.8+	7	3.0+
4	24	21.2	25	20.4	23	14.2+	13	14.5	9	8.3	10	8.5
(cont)												
-4	24	24.0	25	25.5	23	18.1	19	19.4	12	10.8	10	11.0
-5/(=-5	24	13.0+	20	14.2	11	7.2	14	8.2+	11	4.0+	8	4.1+
-6/(=-6	11	5.8+	12	6.6+	9	2.8+	5	3.4	-	-	-	-
-7/(=-7	6	2.3+	6	2.9+	-	-	-	-	-	-	-	-
Chi Square	46.8***		32.2**		33.5***		23.5*		18.9*		13.5	
D. F.	14				12				10			

Note: The "+" in the "E+" column indicates that the contribution of the cell to the total chi square exceeds 3.0; *** p(.001; ** p(.01; and * p(.05.

Table 5
Chi Square Results for Cross-year Within Grade Analysis
Reading and Math Treated Jointly, CRIT=1

	Grade 1		Grade 2		Grade 3		Grade 4	
S	O	E(CHI2)	O	E(CHI2)	O	E(CHI2)	O	E(CHI2)
4	15	2.2(73)	9	2.0(25)	12	2.2(44)	5	1.6(7.4)
3	16	7.9(8.2)	16	9.4(4.6)	11	8.1	20	7.3
2	34	41.8	40	42.6	38	42.0	37	39.7
1	52	70.8	72	83.5	52	74.1	57	77.6(5.5)
0	172	156.7	151	151.7	166	158.7	181	172.3
-1	53	80.8	80	85.8	91	85.7	67	80.7
-2	49	54.9	34	43.9	41	47.7	42	42.0
-3	27	12.0(19)	21	9.9(13)	12	10.4	17	8.0(10)
-4	13	3.9(21)	8	2.1(17)	8	2.5(12)	5	1.8(5.6)
Chi Square(8)		139.6		63.1		66.3		53.7
Cons Chi Sq(6)		97.5***		44.9***		28.2***		53.0***

Note: CHI2 is the contribution to chi square and is included only if it exceeds 3.0; *** indicates p < .001.